The Key Approach to Translation: Word Alignment Models

Timothy Liu
University of California at Berkeley
CS294-19
timothytliu@berkeley.edu

ABSTRACT

This paper focuses on a key aspect of Statistical Machine Translation: word alignment. Various word alignment models are presented, first differentiating between methods and then highlighting the preferred method. A partially detailed mathematical explanation is provided for each model as well as a brief implementation of the Expectation Maximization Algorithm (EM Algorithm) for later models. Furthermore, statistical and error analysis follow each segment of models. The purpose of this paper is to show an integral sub problem that Statistical Machine Translation must deal with and how some computational linguists and computer scientists go about doing it.

General Terms

EM Algorithm, Statistical Machine Translation (SMT)

Keywords

Recall, Precision, Alignment Error Rate (AER)

1. INTRODUCTION

Machine Translation is a method of using software to translate speech or text from one natural language to another. Statistical Machine Translation uses translation methods based on statistical models over traditional models to improve the efficacy of translation. Statistical Machine Translation has the advantage of processing machine-readable text, tailoring translation systems to all pairs of languages, and more natural translations. SMT is centered around the probability of English text being a translation of French text for lack of better example languages. Furthermore, to decode a translation,the best English word associated with a French word must be achievable.

$$p(e|f) \propto p(f|e)p(e)$$

$$\tilde{e} = argmax_{e \in e^*}\ p(e|f) = argmax_{e \in e^*}\ p(f|e)p(e)$$

Word alignment is a natural language processing task of identifying translational relationships among the words, which serve as the foundation for most approaches to SMT. There is a hidden variable in statistical translation models, which is the alignment between the parallel text.

$$Pr(f|e) = \sum_{a_i} Pr(f, a_i|e)$$

Three units of measurement are often used to characterize the proficiency of word alignment models. [1]

[1]http://acl.ldc.upenn.edu/J/J03/J03-1002.pdf

Precision - The fraction of test alignments produced by the method under test that are a part of the possible alignments.

$$Precision = \frac{|A \bigcap P|}{|A|}$$

Recall - The fraction of test alignments produced by the method under test that are a part of the correct alignments.

$$Recall = \frac{|A \bigcap S|}{|S|}$$

Alignment Error Rate (AER) - Given a word alignment model, the fraction of test alignments produced by the method under test that do not correspond to either possible or correct alignments out of all possible and correct alignments.

$$AER = 1 - \frac{|A \bigcap P| + |A \bigcap S|}{|A| + |S|}$$

In these equations, S denotes the set of alignments annotated as *sure*, P denotes the set of alignments annotated as *possible* or *sure*, and A denotes the set of alignments produced by the method under test.

2. HEURISTIC BASED ALIGNMENT MODELS

Most word alignment models use a training corpus consisting of sentence pairs that have correct and possible alignments of words between two natural languages. To train word alignment models, it is necessary to produce some sort of association between one natural language and the other. One method requires an accurate translation probability model.

However, heuristic based alignment models use statistical measures from the training corpus and run a heuristic on the extracted statistics. A heuristic is used to generate associations between pairs of words from parallel corpora to score and evaluate the alignment of sentence pairs. Instead of an accurate probability distribution, only accurate and powerful heuristics are needed.

2.1 Dice Coefficient and Competitive Linking

The Dice Coefficient Model uses a simple heuristic. First it counts up the co-occurence of each English and French word. Then a score is generated for each pair to accurately represent its relation to other scores. The Dice Model uses the simple heuristic of choosing the best English position or alignment with the best score.

Table 1: Alignment Model Test Results

Model	Precision	Recall	AER	# of Extra Sent.
Baseline	.366	.226	.686	0-1,000,000
Dice	.233	.218	.772	0
—	.368	.413	.616	10,000
—	.480	.573	.488	100,000

Table 2: Competitive Linking Test Results

Model	Precision	Recall	AER	Extra Sent.
Competitive	.467	.545	.504	0
Linking	.537	.632	.428	1,000
—	.567	.650	.402	10,000
—	.601	.686	.361	100,000

This process is denoted by these equations:

$$dice(i,j) = \frac{2 \cdot C(e_i, f_j)}{C(e_i) \cdot C(f_j)}$$

$$a_j = argmax_i\{dice(i,j)\}$$

Compared to the baseline word alignment model, it seems reasonable to assume that a better representation of translation would occur with a larger corpus. The heuristic models will generally be more accurate as the error rate decreases and overtakes the baseline model. See **Table 1**.

A better heuristic based on co-occurence counts is the Competitive Linking Alignment Model. First, a more descriptive statistic, G^2, is generated using modified co-occurence counts.

Co-occurence Counts	**e**	**¬ e**
f	a	b
¬ f	c	d

$$B(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$p_1 = \frac{a}{a+b}; p_2 = \frac{c}{c+d}; p = \frac{a+c}{a+b+c+d}$$

$$G^2(e,f) = -2 \log \frac{B(a|a+b,p_1)B(c|c+d,p_2)}{B(a|a+b,p)B(c|c+d,p)}$$

After generating these scores, the heuristic first sorts the scores and then picks the alignment with the best score. Afterwards, the heuristic eliminates all remaining alignments that share a position in the chosen alignment. This creates a better representation of translation as alignments should not be limited to taking the max based on one side of the language. [2] See **Table 2**

2.2 Alignment Matrices and Error Analysis

Alignment Matrices are visual representations of alignments between two natural languages in parallel text. In the matrices, the French words on the y-axis are aligned with the

Figure 1: Dice Coefficient & CL Alignment Matrix

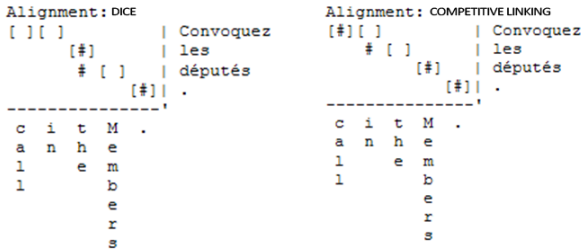
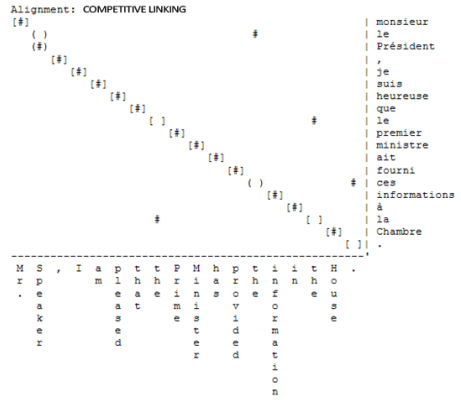


Figure 2: Competitive Linking(CL) Alignment Matrices



English words on the x-axis. The hashes, #, represent proposed alignment pairs from the model while the brackets, *[]*, indicate the actual reference pairs, and the parentheses, *()*, represent possible pairs. See **Figures 1 and 2**.

The Dice Model does well overall in comparison to the baseline; however, there are flaws. The Dice Alignment Matrix shown in **Figure 1** has one such error in that if a single alignment generates a high probability, then that alignment will overshadow every other alignment. This is due to the co-occurence counts not accurately representing the true counts and having too much flexibility in choosing alignments. Therefore, a single word that appears more will often repeat like the English letter "the". See **Figure 1**

However, the Competitive Linking Model fixes this error trend in that it regulates a specific alignment pair so that there is no overlap of alignment on the horizontal and the vertical. This is a little too rigid because some French words require two English words or more and vice versa. However, out of the alignments where no positions repeat, the competitive linking does quite well. See **Figures 1 and 2**

As seen in **Table 1 and 2**, the *AER* steadily decreases, while *precision* and *recall* increase. However, the models are still too simple. One large aspect of this is that the models only do a naive count of the data, which is not an accurate representation. Also, trends like a strong correlation of sure or possible alignments along the diagonals are not exploited.

[2] http://www.cs.nyu.edu/~melamed/ftp/papers/clmote.pdf

3. PROBABILISTIC BASED ALIGNMENT MODELS

Probabilistic Alignment Models lie at the heart of SMT as it is the most comprehensive in estimating one of the three key problems of translation. The three computable problems can be seen in the quantity that best characterizes SMT when using Bayes' Rule.

$$Pr(e|f) = \frac{Pr(e)Pr(f|e)}{Pr(f)}$$

This gives us the first two problems: the *language model probability* and the *translation model probability*. For fear of redundancy, this paper ignores the discussion on language models, which can be discussed at a later time. The last problem is how to use the previous problems' probabilities to determine what the best English string would produce the best overall probability or translation. The foundation for this problem is formalized as a search problem.

Probabilistic Alignment models accurately represent the translation probability best by adding the hidden variable of alignment and by using the Expectation-Maximization Algorithm to do parameter estimation.

Two models are presented to introduce the concept of basic probabilistic alignment models: IBM Model 1 and IBM Model 2. [3]

The main equation used by both is interpreting the Probability of French text given English Text:

$$Pr(f, a|e) = Pr(m|e) \prod_{j=1}^{m} Pr(a_j|a_i^{j-1}, f_1^{j-1}, m, e) Pr(f_j|a_i^j, f_1^{j-1}, m, e)$$

3.1 IBM Model 1

The first model presented as such is very simple and several assumptions are made. First, $Pr(m|e)$ is approximated as a small constant ϵ because the probability of the length of a French text is independent of both the English word and the length. Second, the alignment probability is a uniform distribution with respect to the length of the English sentence.

This is due to the lack of any special order and so all alignments in the English are equally likely. However, when decoding, it is necessary to assume that the $NULL$ position has a small probability(.2) and the other alignment probabilities are distributed uniformly among the remaining(.8).

The last assumption is that translation of a French word given other French words, position, and English word only depends on the English word. Therefore $Pr(f|a, f, e)$ simply becomes the translation probability $t(f|e)$.

$$Pr(f, a|e) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_i|e_{a_j}) \cdot$$

Now the EM algorithm is used to maximize $Pr(f|e)$ by finding the constrained maximum for each $t(f|e)$.

$$c(f|e) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i)$$

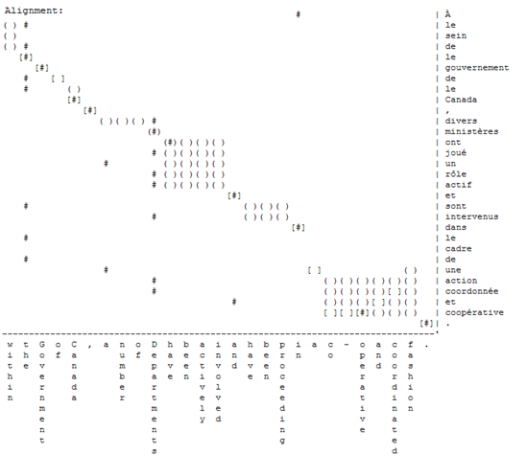
$$t(f|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})$$

Essentially the counts are a distribution of the transitional probabilities not normalized. Each count is changed slightly by the transitional probabilities and then when normalized become the new transitional probabilities. Therefore, after repeating many times, the transitional probabilities converge to an absolute maximum making optimal values of $t(f|e)$. The optimal number of iterations was 40 for 100,000 sentences due to minimal expected return based on time of calculation.

Table 3: IBM Model 1 Test Results

Model	Precision	Recall	AER	Extra Sent.
Model 1	.485	.600	.473	0
—	.526	.676	.429	1,000
—	.554	.738	.376	10,000
—	.624	.781	.311	100,000

Figure 3: IBM Model 1 Alignment Matrix



3.2 IBM Model 2

IBM Model 2 corrects the assumption that the probability of the alignment is equally likely for all alignments. IBM Model 2 prefers alignments that are along the diagonals of the matrix. This follows the correlation seen in the alignment matrices as a descriptor of translation. For this to occur, the alignment probability is now a function of the length ratio between the sentences and the paired positions.

$$Pr(a_j|a_i^{j-1}, f_1^{j-1}, m, e) = \frac{1}{Z} e^{-\alpha(i - j\frac{I}{J})}$$

Because of this new varying probability, there is an extra parameter to train or to validate, *alpha*. Alpha was trained on the test data with 0 extra sentences to achieve an optimal value of .2. This was done by simply testing different alpha

[3] acl.ldc.upenn.edu/J/J93/J93-2003.pdf

Figure 4: IBM Model 1 Alignment Matrices

Table 4: IBM Model 2 Validation

Alpha	Precision	Recall	AER	Extra Sent.
1	.560	.556	.441	0
.8	.566	.577	.430	0
.6	.581	.612	.408	0
.4	.607	.663	.374	0
.3	.618	.663	.373	0
.2	.626	.676	.356	0
.1	.614	.545	.361	0

values for simplicity and efficiency. See **Table 2** .A further aspect of model 2 is that convergence will be to a maximum; however, it may only be a local maximum and not an absolute maximum. Therefore, also for simplicity and efficiency the number of iterations was kept at 40.

Again to decode the null word, a special probability is kept at .2 while the other alignments are distributed among the remaining of .8. Therefore, both IBM Model 1 and 2 are fairly similar with Model 1 being a special case where the alignments are simply based on the length of the English sentence. Model 1 and especially Model 2 perform considerably better as there are more parameters to tune and a better representation model. Analysis of results and errors occur next.

4. ANALYSIS OF RESULTS AND ERRORS

Both Model 1 and Model 2 did considerably better than the other models when looking at over a 100,000 size corpus. See **Table 3 and 5**. The general trend is that a larger corpus produces lower AER and higher *precision* and *recall*. This occurs because all models look at some sort of count of each English word versus the French word. Therefore, with a larger corpus, the more accurate representation the model can attain of the relationship between certain English and French words. See **Figure 7** However, a closer look at the alignment matrices is required to truly evaluate the results from Model 1 and Model 2.

4.1 Alignment Matrices and Error Trends

The alignment matrices for IBM Model 1 show a few error trends. See **Figure 3 and 4**. The major error is the column of alignments that are aligned to the word “departments”, “overproduction”, and “homeowner”. This occurs because Model 1 likes to align French words to rare English words.

Table 5: IBM Model 2 Test Results

Alpha	Precision	Recall	AER	Extra Sent.
.2	.626	.676	.356	0
—	.654	.713	.337	1,000
—	.711	.757	.279	10,000
—	.804	.810	.212	100,000

Figure 5: IBM Model 2 Linking Alignment Matrices

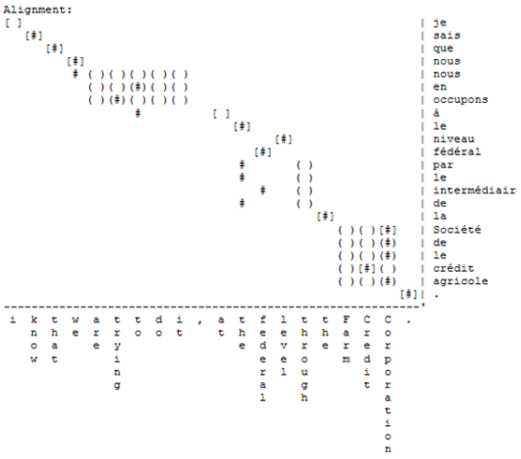


Figure 6: IBM Model 2 Linking Alignment Matrices

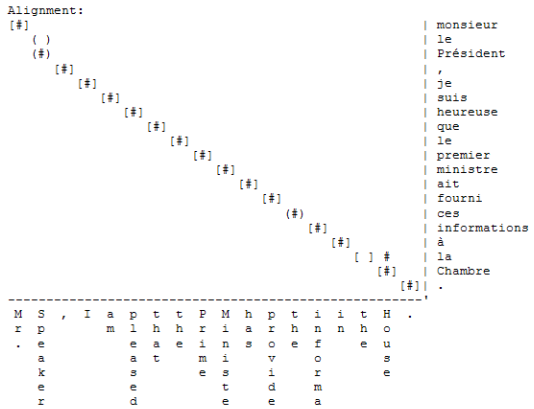
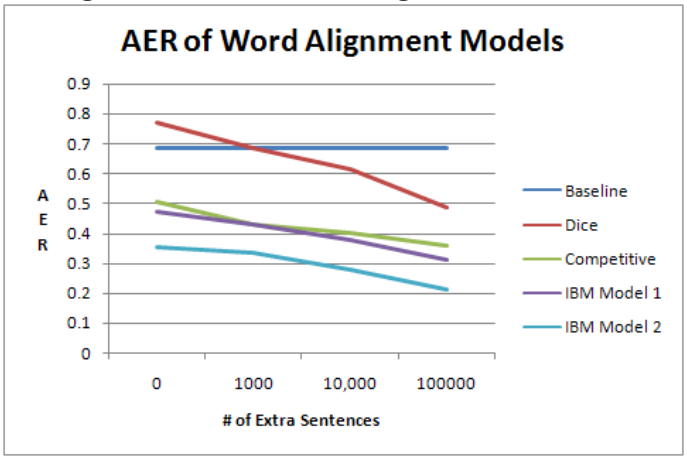


Figure 7: AER of Word Alignment Models



This is due to the low count of e in the denominator of $p(f|e)$ and low co-occurence count, which produces a high ratio. A lesser consequence of this probability is that if the co-occurrence count of pairs of words, "the" and "le", is high and the denominator is high the ratio is also high. Therefore, the probabilities are not representative of how the words should actually be translated. A solution would be to prefer and restrict single alignments.

This is somewhat related to the fact that all of these models only train with single word associations(one-to-one). Although many-to-one alignments appear, it is not truly representative of actual translations and problems like aligning words randomly to rare words occur. Therefore, true phrases are only aligned properly some of the time and one-to-many can never be aligned as $a_j = i$ for all word alignment models. The competitive linking model restricted the models to one-to-one; however, in **Figure 4** two correct alignments occur in the same column. Also, many of the possible alignments in parenthesis () are often blank. Therefore, there is also a tension in how flexible alignments can be.

In Model 2, a successful advantage is the alignment along the diagonals; however errors still occur in other cases. See **Figure 5 and 6**. Essentially all the phrase based alignment errors remain in Model 2 as they occurred in Model 1 since they are extremely similar. Therefore, the general analysis is that rare English words have inflated probabilities. There is a distinct lack of handling phrasal alignment and restricting alignments is a delicate process.

5. CONCLUSION

Therefore, Word Alignment Models create excellent associations between two parallel corpora, which is extremely useful for determining translation probability. This is useful because these alignment models can be used in any context because they are not based on grammar rules or any other language specific characteristic. Translation probability and its key component, word alignment, is integral to Statistical Machine Translation.

[1] Franz Joseph Och and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." Computational Linguistics.

[2] I. Dan Melamed. 2000. "Models of Translational Equivalence among Words." Computational Linguistics.

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." Computational Linguistics.